Generative Representational Instruction Tuning

Year: 2024 | Citations: 198 | Authors: Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei

Abstract

All text-based language problems can be reduced to either generation or embedding. Current models only perform well at one or the other. We introduce generative representational instruction tuning (GRIT) whereby a large language model is trained to handle both generative and embedding tasks by distinguishing between them through instructions. Compared to other open models, our resulting GritLM 7B sets a new state of the art on the Massive Text Embedding Benchmark (MTEB) and outperforms all models up to its size on a range of generative tasks. By scaling up further, GritLM 8x7B outperforms all open generative language models that we tried while still being among the best embedding models. Notably, we find that GRIT matches training on only generative or embedding data, thus we can unify both at no performance loss. Among other benefits, the unification via GRIT speeds up Retrieval-Augmented Generation (RAG) by>60% for long documents, by no longer requiring separate retrieval and generation models. Models, code, etc. are freely available at https://github.com/ContextualAI/gritlm.